J. Mol. Biol. (1981) **153**, 1027–1042

Comparative Model-building of the Mammalian Serine Proteases

JONATHAN GREER

Department of Biological Sciences
Columbia University, New York, N.Y. 10027, U.S.A.

(Received 4 September 1980, and in revised form 7 July 1981)

Proteins have been classified into families based upon sequence homology. An accurate, systematic comparative model-building procedure for a homologous family of proteins would be very valuable scientifically. This paper presents such a procedure and applies it to the mammalian serine proteases, which are ubiquitous and involved in many important biological functions. Eleven proteins of this family are considered here, including a variety of blood serum, intestinal and pancreatic proteins as well as a closely related bacterial enzyme.

The modeling method capitalizes upon the availability of three experimentally determined structures for mammalian serine proteases. These structures show that the molecule is divided into structurally conserved regions, which contain the strong sequence homology, and structurally variable regions, which include all the additions and deletions. We show that by applying this structural distinction to new sequences, erroneous alignments of the sequences are greatly minimized.

For each aligned new sequence, the structurally conserved regions can be constructed from any of the known structures. In examining the variable regions, we have found that a variable region that has the same length and residue character in two different known structures usually has the same conformation in both. Thus, when the eight structurally unknown proteins are modeled, most of the variable regions can be constructed directly from the known structures. A minority of the variable regions require more sophisticated analysis to evaluate the relative merits of a small number of possible conformations. Only a very few are so different that modeling by homology is entirely ruled out. We demonstrate, therefore, that by this modeling procedure, the maximum of each of these mammalian serine proteases is constructed directly from the experimentally determined structures and the necessity to build from intuition or from energy considerations is greatly reduced.

1. Introduction

Many diverse proteins have been classified into families based upon sequence homology (Dayhoff, 1972). The similarity of three-dimensional structure in these homologous families has suggested that if the structure of one of these proteins is known, then it should be possible to construct the three-dimensional structure of other members of the same family by comparative model building.

Brown *et al.* (1969) first applied this method to the construction of α-lactalbumin from the homologous lysozyme structure. Hartley (1970) used it to identify the functionally important residues in the chymotrypsin-like serine protease family.

0022-2836/81/361027-16 $02.00/0

Jurasek *et al.* (1976) have built a structure for *Streptomyces* trypsin-like protein from that of bovine trypsin. In the most ambitious effort, McLachlan & Shotton (1971) built a model of α-lytic protease, a bacterial serine protease, from the elastase structure (Shotton & Watson, 1970), a mammalian serine protease, even though the sequence homology was quite weak. The X-ray structure of this protein was subsequently solved by Delbaere *et al.* (1979) and showed significant differences from the model structure. Extrapolating even further, Kretsinger (1976) argued that a variety of Ca^{2+} binding proteins such as troponin C, myosin light chain, and others are structurally homologous to carp Ca^{2+}-binding protein.

Successful comparative model building depends upon how closely the structure that one is attempting to build fits the known structure. With our present state of understanding of protein structure, the only measure that can be applied is examination of the extent of sequence homology between the known and unknown proteins. The conclusion of the comparative studies cited above is that structural homology persists even when sequence homology is hardly detectable. However, for the purpose of comparative model building, the reverse is important, i.e. the presence of sequence homology is necessary to indicate structural homology. Thus, the first step in comparative model building is the alignment of the new sequence with that of the known structure.

Model building is usually applied to proteins that have significant numbers of relative additions and deletions in their primary structures. A variety of methods have been developed to align such protein sequences and to measure the degree of sequence homology. These include minimum base distance (Jukes & Cantor, 1969; Dayhoff, 1972; Fitch, 1966) and application of observed amino acid substitution frequencies (McLachlan, 1971). Such methods have been used successfully to discover related proteins and to construct evolutionary trees (Dayhoff, 1972; De Haen *et al.* 1975), and to recognize internal gene duplication (McLachlan, 1972).

For model building, however, it is not sufficient to obtain a general relatedness of two sequences. In order to construct a new structure correctly, it is necessary to align the two sequences accurately and unambiguously at *every* residue where the two structures are homologous. In cases where the two proteins are sufficiently diverged and several additions and deletions occur, the established methods of sequence alignment by maximizing sequence equivalence and homology will also align residues that are equivalent only by chance but do not correspond in structure. Any incorrect alignment of this kind guarantees that the built structure will be incorrect at this site. A prime example of this was the alignment by McLachlan & Shotton (1971) of Cys137 and 159 of α-lytic protease with the disulfide bridge between Cys168 and 182 of the methionine loop in elastase. Actually, Delbaere *et al.* (1979) have shown that the structures of these two cysteines are very different in the bacterial enzyme. Thus, it is essential for model building to take into account all available information about the sequence and the structure to produce the optimum alignment of the two sequences at every possible residue and also to recognize where alignment may be inappropriate.

Examination of the model-built and experimental structures for α-lytic protease indicates that comparative model building from the mammalian enzymes may be satisfactory for some parts of this molecule, but that it is inadequate and even

misleading in the novel parts of the molecule because of the weak homology. The correct construction of these structurally novel regions cannot be accomplished by extrapolation from the mammalian serine proteases using comparative model building. It requires knowledge of protein structure, folding, and energetics, which are unfortunately beyond our current capabilities. Therefore, despite the apparent close relationship of the bacterial serine proteases to the mammalian enzymes, the problem of constructing the bacterial enzymes is not treated in this paper. Instead, the more conservative, yet still demanding, task of developing systematic methods for constructing structures of new members within the closely related mammalian serine protease family is examined.

Serine proteases are ubiquitous in mammalian tissue and are involved in a variety of important functional roles such as blood coagulation and complement fixation. In this paper we describe a new sequence alignment and modeling procedure, which capitalizes upon the availability of several known structures of mammalian serine proteases and the observed homology between their sequences. This modeling method is applied to a wide variety of mammalian serine protease sequences that are available. A simpler form of this method was previously employed to construct a model of the human serum protein, haptoglobin heavy chain (Greer, 1980). In the accompanying paper, the procedure is applied in detail to a particular case, i.e. the structure of blood clotting factor X_a and the activation peptide of prothrombin. These model structures provide new insight into the nature of the highly specific interaction between these two proteins.

2. Method of Model Building

(a) *Structural features of the mammalian serine protease family*

One would expect the structure of any particular protein of the mammalian serine protease family to consist of parts that are common to all the members of this family and parts that are idiosyncratic to that protein. The common parts should form the structural framework and provide the common functional properties, while the protein specific parts should reflect the individual properties of the respective protein. If the regions of the molecule that are common can be identified, then they may be used as the basic framework for the construction of atomic co-ordinates for any of the proteins of the family.

It is fortunate that 3 independently determined structures of mammalian serine proteases are available at atomic resolution: chymotrypsin (Birktoft & Blow, 1972; Tulinsky *et al.*, 1973), trypsin (Huber *et al.*, 1974; Stroud *et al.*, 1974) and elastase (Shotton & Watson, 1970; Sawyer *et al.*, 1978). These structures were aligned relative to each other by least-squares fitting of the α-carbons as previously described (Greer, 1980) (see Fig. 1). A residue-by-residue correspondence was made using the 3-dimensional superposition of the α-carbons. This correspondence is summarized by aligning the related residue names along the 3 sequences (Fig. 2). Examination of the 3-dimensional structures (Fig. 1) shows that large parts of the 3 molecules are closely similar in structure and hence appear to be structurally conserved, while other sections differ considerably and are quite variable. Groups of residues (and not just single α-carbons), where the α-carbons of all 3 structures lie very close to each other (the *maximum* deviation permitted for any α-carbon in the group is 1 Å), are designated structurally conserved regions or SCRs. These are enclosed in boxes in Fig. 2. Variable regions or VRs are those residues where deviation of virtually every α-carbon exceeds 1 Å. Careful analysis shows that the SCRs are the β-barrels, β-sheet and α-helix. Each distinct structural element, i.e. β-strand or α-helix, is placed in a separate

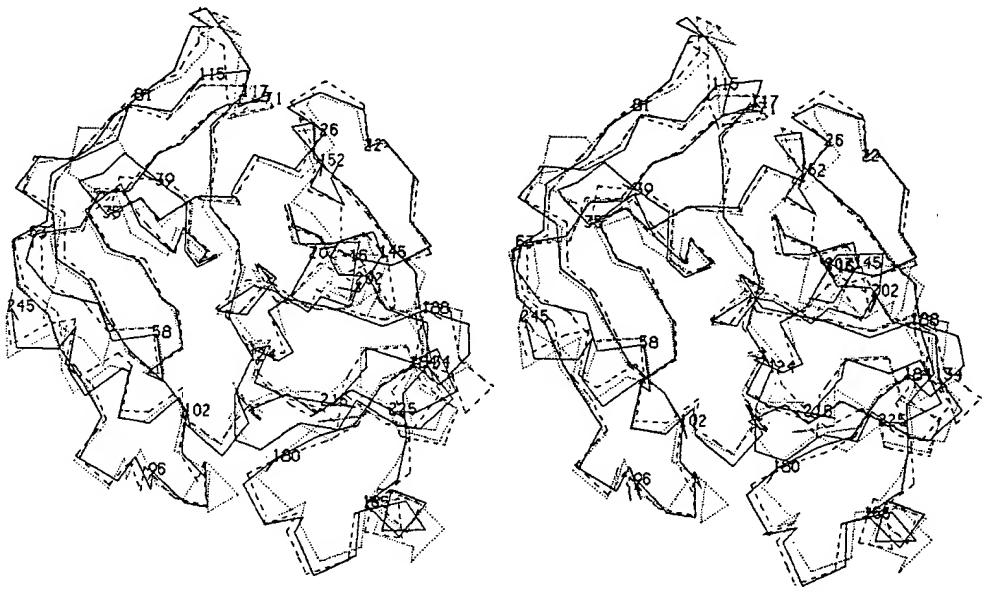


FIG. 1. Stereo presentation of α-carbon plots of chymotrypsin (solid lines), trypsin (broken lines), and elastase (dotted lines). All 3 molecules have been placed in the chymotrypsin co-ordinate frame by least-squares fitting as previously described (Greer, 1980). The residue numbers are those of chymotrypsinogen. The residue labels mark the borderlines between the SCRs and the VRs and are the residues in the SCR just before and just after a VR (see Fig. 2). The view of the molecules is looking into the serine protease active site with the specificity pocket positioned to the right.

box. Thus, the box encompassing residues 39 to 58† is divided in two because it consists of 2 β-strands. A variable region usually corresponds to one external loop in the molecule. This is exactly where variation in the structure is to be expected.

Fig. 2 demonstrates that each SCR shows strong sequence homology, while the VRs show little sequence homology and are the sites of addition and deletion of residues. It is entirely reasonable that SCRs should have identical or closely homologous sequences, since these residues are performing the same structural or functional role in each of these proteins. This reasoning requires that the structural conservation, which was chosen above solely on the basis of the closeness of α-carbon position, also applies to the side-chain positions. Detailed examination of the 3 structures shows that, in most cases, the side-chain positions agree very closely as well. The exceptions are noted by asterisks in Fig. 2. They always involve residues on the external surface of the protein as judged by accessibility calculations (Lee & Richards, 1971; Richards, 1977).

(b) *New method of alignment of the mammalian serine protease sequences*

A large number of sequences are available for proteins of the mammalian serine protease family, from a variety of different sources and species (Dayhoff, 1972,1978). For this study, one sequence for each protein was selected. In each case, only the part of the protein that corresponds to the serine protease sequences of Fig. 2 is considered here. The proteins and

† The residue numbering used throughout corresponds to the standard chymotrypsinogen residue numbering (Hartley, 1970).

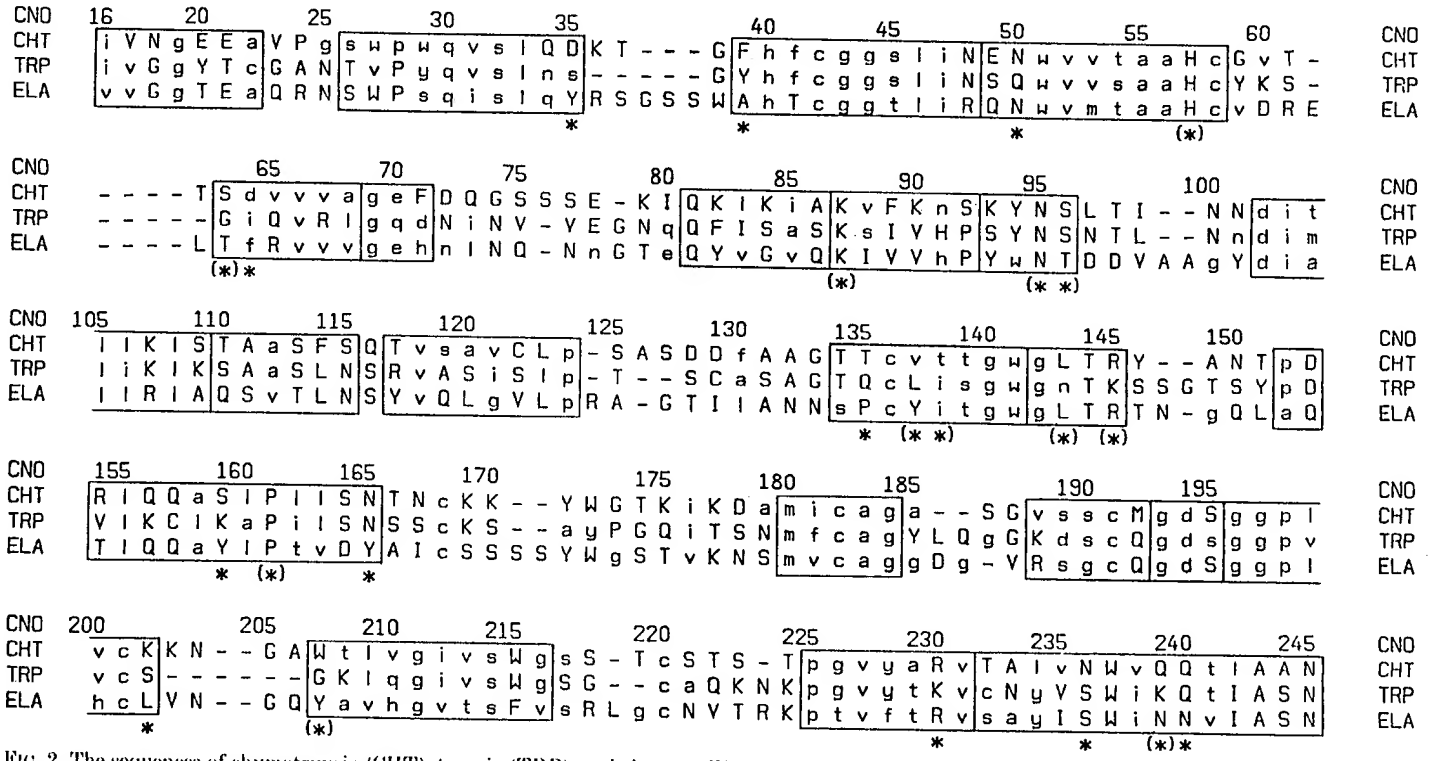


FIG. 2. The sequences of chymotrypsin (CHT), trypsin (TRP), and elastase (ELA) are aligned by comparison of the 3-dimensional structures (see text). The top row, labeled CNO, gives the standard chymotrypsinogen numbering. The boxes indicate the structurally conserved regions (SCRs) in these three structures based upon α-carbon position (see Fig. 1). Residues outside the boxes are the variable regions (VRs). The asterisks mark the residues in the SCRs where the side-chain positions differ markedly among the 3 structures. An asterisk in parenthesis means the side-chain positions differ, but less markedly. Residue names are given in the standard one letter code: a, A = Ala; b, B = Asx; c, C = Cys; d, D = Asp; e, E = Glu; f, F = Phe; g, G = Gly; h, H = His; i, I = Ile; k, K = Lys; l, L = Leu; m, M = Met; n, N = Asn; p, P = Pro; q, Q = Gln; r, R = Arg; s, S = Ser; t, T = Thr; v, V = Val; w, W = Trp; y, Y = Tyr. A residue name is written in upper case letters if its side-chain is accessible to solvent in the 3-dimensional structure. Solvent accessibility (Lee & Richards, 1971; Richards, 1977) was calculated using a program written by Dr Bruce Bush. Residues which showed more than 20 Å^2 accessible surface area in their side-chain atoms were taken as accessible. Less than 20 Å^2 was taken as buried and is represented in lower case.

TABLE 1

Serine protease family proteins of known sequence

Code	Protein	Species	Source	Number of residues	Reference
CHT	Chymotrypsin	Bovine	Intestine	228	Birktoft & Blow (1972)‡
TRP	Trypsin	Bovine	Intestine	223	Huber *et al.* (1974)‡
ELA	Elastase	Porcine	Intestine	240	Shotton & Watson (1970)‡
HPH	Haptoglobin heavy chain	Human	Blood	245	Kurosky *et al.* (1980)
KAL	Kallikrein	Porcine	Pancreas	192†	Tschesche *et al.* (1976)
FIX	Factor IX_a (Christmas factor)	Bovine	Blood	235	Katayama *et al.* (1979)
FAX	Factor X_a (Stuart factor)	Bovine	Blood	232	Titani *et al.* (1975)
PLM	Plasmin B chain	Human	Blood	230	Wiman (1977)
GSP	Group-specific protease	Rat	Intestine	224	Woodbury *et al.* (1978)
THR	Thrombin B chain	Bovine	Blood	257	Magnusson *et al.* (1975)
SGT	Bacterial trypsin	*S. griseus*		221	Olafson *et al.* (1975)

† This is an incomplete sequence with a part missing from the center of the protein.

‡ These references refer to the 3-dimensional structures. The sequences were obtained from Dayhoff (1972) in these cases with the original sources cited therein.

sequences used are listed in Table 1 and include a variety of blood, intestinal and pancreatic proteins.

The alignment method, with its novel aspects, is as follows.

(1) Alignment by sequence homology is limited to stretches of clear and unequivocal sequence homology. Each such section of the new sequence is thereby related to its respective SCR.

(2) The remaining positions in the SCRs are filled sequentially, without permitting additions or deletions within that SCR.

(3) The VRs are aligned arbitrarily at this stage unless there is significant sequence homology (which occurs very rarely). A tentative alignment is considered when a half-cystine appears, which is characteristic of that VR. Final alignment will be performed during the modeling process when the VR as a whole is compared to those of the known structures as will be described later.

Step (1) was performed by a simplified form of the methods of McLachlan (1971,1972). In fact, this alignment could equally well be performed by hand. Comparison matrices were generated with a span length of 5 residues using equal weights of 1. The results obtained were relatively insensitive to the exact parameters used because the stretches of sequence homology that occur in the SCRs are usually so clear. Since the emphasis was upon structural conservation, homologous residues were chosen to be those with similar structural properties such as size, charge, polarity and hydrophobicity (see Table 2). Minimum mutation distance (Fitch, 1966; Jukes & Cantor, 1969; Dayhoff, 1972) or observed amino acid substitution frequencies (McLachlan, 1971) were not used to determine homology, as they can be expected to include a variety of other factors in addition to structural equivalence. Residue identities were scored as 1·0, while homologies (Table 2) were scored as 0·5. Gaps were counted as 0. Each new sequence was compared with the 3 standard sequences, chymotrypsin, trypsin and elastase, as aligned in Fig. 2. Each of the standard sequences was also examined by comparing it to the remaining 2 standard sequences. The calculated comparison values were normalized on a per residue basis resulting in values

TABLE 2

Homologous residues

Set	Residues
1	D E K R
2	G A V
3	A V L I
4	V L I M
5	F Y W
6	S T
7	Q N
8	G P (for turns)

between 0 (meaning no similarity) and 1 (meaning perfect identity for all 5 residues of the scan with all the standard sequences). Values of 0·5 or greater were taken as significant sequence homology in this work. The full comparison matrices that were calculated are too long to be included here and are similar to those described by McLachlan (1971,1972).

(c) *Modeling a new structure*

The new sequence, which has been aligned by the above procedure, has been parsed into separate structural elements: SCRs and VRs. This division of the structure into semi-independent SCRs and VRs allows separation of the modeling problem into 2 distinct parts.

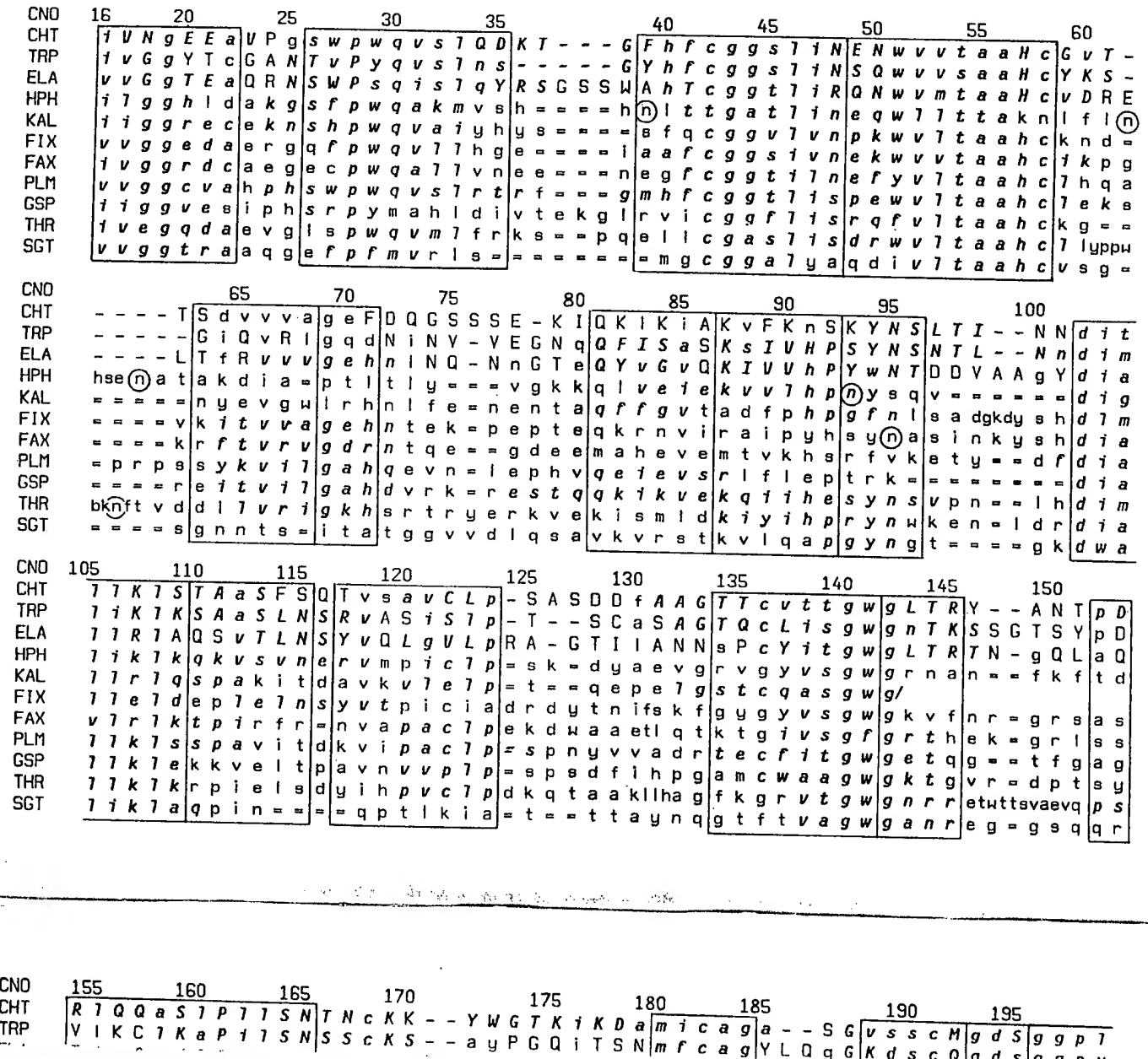
A model for the SCRs of a new sequence is constructed in a relatively straightforward manner from the atomic co-ordinates for the known structures. For a particular SCR, any one of the known structures can be used to model the co-ordinates of the main chain. Co-ordinates of identical side-chains are used directly. For different side-chains, co-ordinates are usually built in a conformation similar to the corresponding side-chain in one of the known structures.

Constructing the VRs for a new sequence is a more challenging task. The goal is to construct as much of the structure as possible, either directly from or by analogy to the experimentally determined known structures and to identify those regions that cannot be built without more sophisticated analysis.

A detailed examination of the VRs in the known structures shows that in many cases, when a particular VR has the same length and residue character in 2 of the 3 known structures, then their conformation is the same. These loops can be considered structurally conserved subsets of the VRs. They indicate that when a new sequence has the same length and residue character in one of these VRs, then that VR is also a member of this structurally conserved subset, and its structure will be the same as that of the known structures.

When we study the distribution of these structurally conserved subsets of the VRs in the known structures, we find the interesting result that a particular known structure will have the same conformation as another of the known structures in one VR, but the same conformation as a third known structure in another VR. For example, chymotrypsin looks like trypsin and not like elastase at the VR at 97–101, while at 203–206 chymotrypsin is the same as elastase and different from trypsin. Thus, if the chymotrypsin structure were being constructed from the other 2 known structures, trypsin would be the appropriate model for the 97–101 loop and elastase for the 203–206 loop. This indicates that, when building the model of a new protein, it is important to examine all the known structures that are available, since the VRs of a new sequence may be best modeled from parts that are selected from several different known structures.

The model of a new protein is constructed by selecting the most suitable SCRs and VRs from amongst the various known structures, changing side-chains to fit the new sequence as


CNO
CHT
TRP
ELA
HPH
KAL
FIX
FAX
PLM
GSP
THR
SGT
16 20 25 30 35 40 45 50 55 60
65 70 75 80 85 90 95 100
105 110 115 120 125 130 135 140 145 150
155 160 165 170 175 180 185 190 195

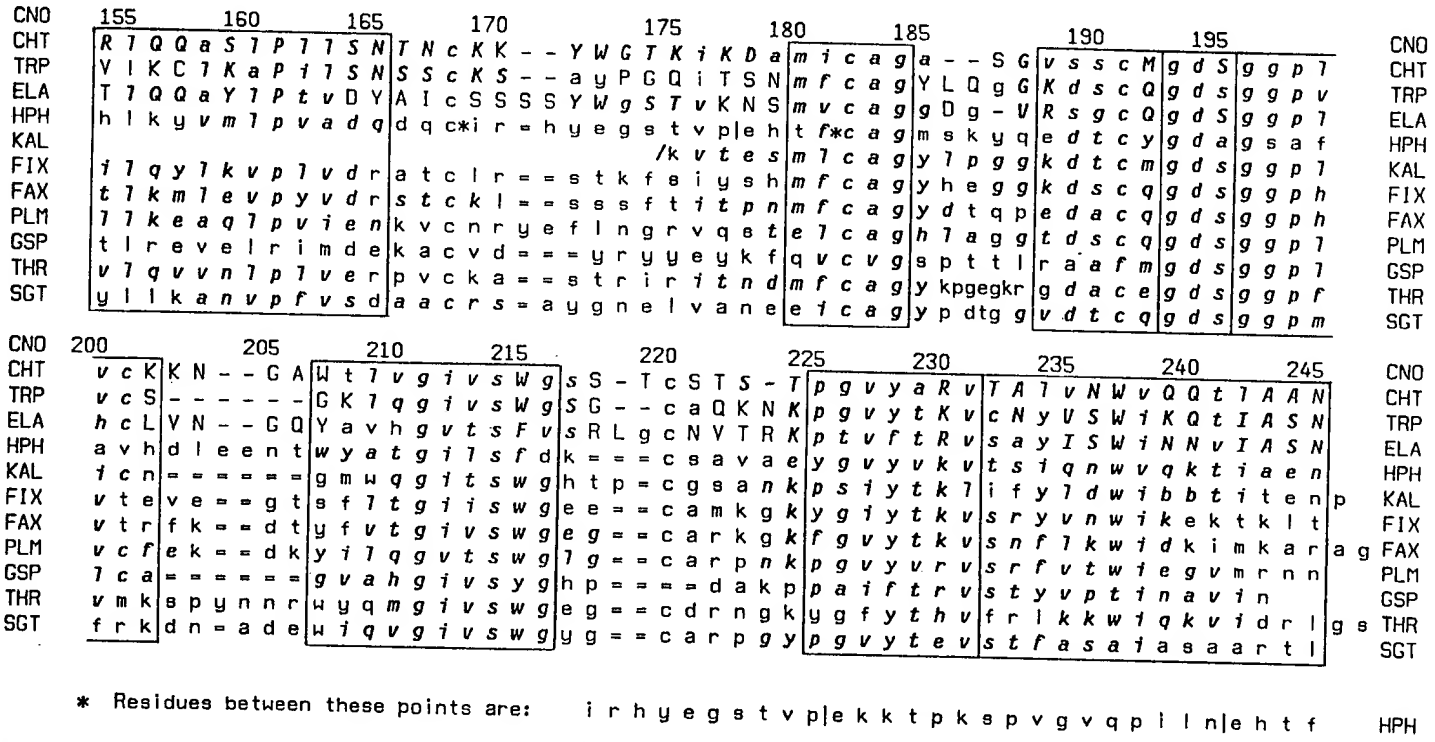


FIG. 3. Aligned sequences for the serine protease family. Residue numbering as in Fig. 2. The first 3 sequences are chymotrypsin, trypsin and elastase, aligned from the known 3-dimensional structures as in Fig. 2. Upper case letters represent solvent accessible residues as in Fig. 2. "–" is used to mark gaps in the relative alignment of the known structures. Residue stretches in each sequence that are homologous to the other 2 sequences by the test described in the text are shown in italics. The rest of the lines give the alignments found for the sequences of the other serine protease proteins. Protein names correspond to the abbreviations given in Table I. The italicized residue names show those regions of the molecule that satisfied the homology test given in the text. "=" is used in these sequences to mark gaps in the relative alignments. As the true structures in the VRs are unknown, the gap positions are often arbitrary (see text). The sequence between the slashes in kallikrein has not yet been determined (Tschesche *et al.*, 1976). The sequence between the asterisks in HPH includes a large addition that appears at the bottom of the Figure. (n) represents Asn residues that have carbohydrate bound in the native protein. The boxes delineate the SCRs as in Fig. 2.

necessary. If the structure of a particular VR cannot be deduced by analogy, it can be left out until more experimental data are available or until a proper energy analysis can be performed.

This modeling procedure guarantees that no overlaps occur between main-chain atoms. Overlaps may be caused by side-chain atoms, but they can usually be relieved simply by varying the side-chain dihedral angles. The conformation of the main chain and side-chains may then be "fine-tuned" to optimize packing.

3. Results and Discussion

(a) *Features of the aligned sequences*

The alignments for the proteins in the serine protease family are presented in Figure 3. The top three sequences are those of chymotrypsin, trypsin and elastase aligned from the three-dimensional structures (see Fig. 2). Sequence stretches that were found to be homologous by the homology criteria are shown in italics. The remaining eight sequences in Figure 3 were aligned using the method described above. Residues that gave a homology index of 0·5 or greater are shown in italics.

Virtually all the SCRs have strong homology in every one of the sequences studied. While there is variation between the sequences as to the exact beginning and end of the homologous stretches that are found, they cover much of the SCR in each case. Thus, the correspondence between sequence homology and conserved structure, observed in the known structures (Greer, 1980), is strongly confirmed in this wide variety of sequences. Consequently, locating stretches of strong sequence homology can be used to align the appropriate parts of a new sequence to the SCRs.

We consider it a basic requirement for constructing accurate models by comparative model building that virtually all the SCRs of a new sequence show strong sequence homology. When this correspondence is not found, it becomes impossible to recognize when significant structural deviations do occur in the new protein. Therefore, the model is likely to contain serious errors, as in the prediction of the α-lytic protease structure (McLachlan & Shotton, 1971), as was shown by Delbaere *et al.* (1979) from the experimental structures of the bacterial enzymes.

There are several SCRs where sequence homology falls below the significance level used in this work. This occurs in the SCR stretches at positions 63–71, 81–96 and 110–115, where in each case several sequences fail to show homology (see Fig. 3). It is interesting to note that sequence homology is absent or weak in these SCRs even among the three known structures, chymotrypsin, trypsin and elastase. The likely explanation for the lack of homology in these SCRs is the high percentage of solvent-accessible residues in these SCRs†, which allows greater side-chain variation. The exact alignment of these sequences in the region of these SCRs is more difficult to deduce and in some cases alternative assignments to those in Figure 3 may have to be considered. This is important for modeling, of course.

The VRs show much less homology. The alignment scheme places all additions and deletions in the VRs; hence, the lengths of the VRs are usually different in

†While it does not appear from the accessibility data (see Fig. 2) that most of the residues are accessible in the SCRs at 63–71, several additional positions such as 64, 70 and 71 must be accessible to solvent since some of the side-chains at these positions in the known structures are charged and cannot be buried.

TABLE 3

Number of residues in the variable regions

	23–25	36–38	59–62	72–80	97–101	124–133	146–151	166–179	185–187	203–206	217–224
CHT	3	3	4	9	5	9	6	14	3	4	8
TRP	3	1	3	9	5	7	6	14	5	0	8
ELA	3	6	5	9	7	9	5	16	4	4	10
HPH	3	2	10	7	1	8	4	31	5	6	7
KAL	3	2	3	9	9	7	—	—	5	0	9
FIX	3	2	5	9	7	11	5	14	5	4	8
FAX	3	3	5	8	5	11	5	14	5	4	8
PLM	3	3	8	9	−1	9	4	16	5	4	8
GSP	3	6	3	9	5	9	5	13	5	0	6
THR	3	4	13	10	6	12	11	14	8	6	8
SGT	3	−2	4	10	3	7	5	15	6	5	8

See Table 1 for abbreviations.

each protein. These are summarized in Table 3. In general, there is a remarkable degree of variability in the lengths of the VRs. This will be discussed in more detail below.

(b) *Significance of this alignment procedure*

It is clear that wherever the mammalian serine protease sequences are strongly homologous, any alignment method will give the same result. The important advantages of the method introduced here are evident in the less- and non-homologous regions of the molecule, where incorrect sequence alignments that result from chance sequence identity or homology are avoided. These incorrect alignments would inevitably lead to erroneous structures by model building. In particular, the new method restricts sequence alignment by homology to stretches of residues in the SCRs where the structural conservation implies that such homology should occur. When non-homologous residues are found in an SCR, alignment proceeds within the restrictions implied by structural conservation, i.e. that no additions or deletions can be permitted. The VRs, which are largely non-homologous, are aligned only by comparing the complete VR sequence with those of the known structures, as discussed in the next section.

Significant differences between this alignment method and one based completely upon sequence homology can be demonstrated by applying these two methods to the sequences of the known structures. Two illustrations are given here; many more can be found by comparing the alignment in Figures 2 or 3 with, for example, Alignment 8 of Dayhoff (1978).

Position 207 lies in an SCR, yet is occupied by either a large aromatic side-chain or a glycine residue (see Fig. 3). Using simply sequence homology, these residues would *not* be assigned to the same position (see, e.g. Alignment 8 of Dayoff, 1978 or

Fig. 4 of Jurasek *et al.*, 1976). We can now ask how it comes to be that such different residue types occupy the same site. The sequence alignment in Figure 3 suggests an explanation. The type of residue that appears in position 207 is correlated with the size of the VR loop at positions 203–206. If no residues appear in this loop, as occurs in trypsin, kallikrein and GSP†, then a Gly is found at 207 and forms one of the residues of the β-bend. If, however, four or more residues are found in this loop, as in the remaining sequences, then residue 207 is usually a Trp or a Tyr, which acts as a spacer between this loop and residues at positions 27, 29 and 137. Both these alternatives appear amongst the three known structures (see Figs 1 and 3). Thus, a C_α position that is structurally conserved in the three known structures seems to have two different roles, depending upon the size of the neighboring loop, and consequently, two different types of residues occupy this site.

Another example is taken from the alignment of the VRs in two of the known structures: chymotrypsin (CHT) and trypsin (TRP). Dayhoff (1978) aligns positions 170 to 175 as follows (CNO is standard chymotrypsinogen numbering):

CNO	165		170				175			180
CHT	N	T	N	C	K	K	—	—	Y	W	G	T	K	I	K	D	A	M
TRP	N	S	S	C	K	S	A	—	Y	P	G	—	Q	I	T	S	N	M

The chance homology in chymotrypsin and trypsin of the Tyr and Gly causes the misalignment of residues 171–174. This would lead to a very different and incorrect model for this loop if trypsin were constructed from chymotrypsin. In fact, the known structures of these two proteins in this loop are identical (see Figs 1 and 3) and this is an example where the VRs in chymotrypsin and trypsin are a structurally conserved subset of the known structures for this VR. Hence, this sequence must be aligned based upon the structural homology and not the sequence homology.

(c) Modeling the mammalian serine protease structures

Using the aligned sequences in Figure 3, model structures can be constructed for each of these proteins using the method described in Method of Model Building, section (c).

The SCRs for each protein are constructed from any one of the known structures. Although the main chain co-ordinates for these SCRs will not overlap, the substitution of different side-chains in a new sequence may cause overlap. In every case where this has occurred in the SCRs, rotation about the side-chain χ angles was sufficient to relieve the overlap. One example of this is the collision of a large residue at position 208, Tyr in HPH and thrombin, Phe in FIX and FAX, and Met in kallikrein, with Pro124 when position 208 is modeled after the Thr in chymotrypsin or the Ala in elastase. However, rotation of χ_1 of the residue at 208 to the position of the Lys at 208 in trypsin resolves this overlap entirely. Occasionally,

†Abreviations used: GSP, group-specific protease; HPH, haptoglobin heavy chain; FIX, blood clotting factor IX_a; FAX, blood clotting factor X_a; SGT, *Streptomyces griseus* trypsin-like protein. See Table 1 for details.

TABLE 4

Closest known model structure and case number for each VR†

	23–25	36–38	59–62	72–80	97–101	124–133	146–151	166–179	185–187	203–206	217–224
HPH	4a	2C	5	3a	3ct	3ce	3e	5	1T	2E	3c
KAL	4a	2C	1T	4a	3e	1T	—	—	1T	1T	3e
FIX	4a	2C	1E	4a	1E	3ce	1E	1CT	1T	1CE	1T
FAX	4a	1C	1E	3a	1CT	3ce	1E	1CT	1T	1CE	1T
PLM	4a	1C	3e	4a	3ct	1C	3e	1E	1T	1CE	1T
GSP	4a	1E	1T	4a	1CT	1C	1E	3ct	1T	1T	3ct
THR	4a	2E	5	3a	3e	3ce	5	1CT	3t	2E	1T
SGT	4a	2T	1C	3a	3ct	1T	1E	3e	3t	2E	1T

† The number refers to the case into which this VR falls, see text. Abbreviations are as follows: C, CHT; T, TRP; E, ELA; and "a" stands for all 3 known structures. Upper case letters represent the cases where the known structures can be used directly or by very close analogy to build the respective VR. Lower case letters show the probable closest known structure that may be useful for a starting conformation for that VR. However, it definitely needs to be modified to fit the actual new sequence probably by some energy analysis (see text).

side-chain rotation is required for other conformational reasons, as in FAX where Cys22 and Cys27 need only be rotated about χ_1 to permit the formation of a disulfide bridge that is unique to this protein.

The VRs are constructed upon this framework of SCRs. When the VRs of the proteins of Figure 3 are modeled based upon the known structures, the following cases can be distinguished (Table 4).

(1) *The length of the VR in the new sequence is the same as that of one or more of the known structures and the nature of the side-chains is consistent with this conformation.* The important observation that structurally conserved subsets of the VRs occur when residue length and character are conserved is a very powerful tool for modeling the VR structures. Thus, for a particular VR, one can often choose from amongst the different known structures for the one that is the same in length and in residue character as that of the new sequence. For example, in the VR at 166–179, FIX, FAX and thrombin can be modeled after chymotrypsin or trypsin, while plasmin is probably patterned after elastase. Similarly, in the VR at 97–101, FAX and GSP can be constructed from chymotrypsin or trypsin and FIX can be built from elastase. In each of these examples, the nature of the side-chains is such that the model conformation is reasonable. For example, no charged residues are buried and no small side-chains are replaced by large bulky ones, unless they are on the surface where there is room to accommodate the additional atoms. A striking demonstration of the influence of residue character is the VR at 217–224. Even though both chymotrypsin and trypsin have the same eight-residue length, their conformations are quite different due to the Cys at position 220, which is the fourth residue in the VR in chymotrypsin but the third residue in the VR in trypsin. Therefore, the equivalent length VRs in FIX, FAX, plasmin, thrombin and SGT should all be modeled from trypsin and not from chymotrypsin (Table 4).

(2) *The known structures show that regardless of length, the VR has a common structural motif, which can be readily extended to a new sequence with a VR of different length.* One illustration of this is the VR at 36–38. The known structures (Fig. 1) indicate that this VR forms the turn between two β-strands, 26–35 and 39–48. The observed effect in the known structures of a longer or a shorter VR is to lengthen or shorten, respectively, the extension of the two β-strands into the VR on both sides of the turn. Thus, a reasonable model for this VR can be constructed for a different length VR found in a new sequence using this structural motif of β-strands ending in a turn. Another VR that can be modeled in this way is 203–206, which forms the turn between the β-strands at 196–202 and 207–216 (see Figs 1 and 3 and Tables 3 and 4).

(3) *The VR in the new sequence differs in length by a relative deletion or by a small relative addition from those of the known structures.* This is also a common occurrence. An example is the six-residue VR in thrombin at 97–101, which is intermediate between the five-residue VR in chymotrypsin and trypsin, and the seven-residue VR in elastase. Similarly, the four-residue VR at 146–151 in HPH and plasmin can be modeled from the known structures where the VRs are five or six residues long.

(4) *In some VRs, the conformations found among the known structures vary considerably even though the length of the known VRs is the same.* A prime example of this is the VR at 72–80, where the three known structures of this VR are quite different yet all contain nine residues (see Fig. 1). Similarly, the VR at 23–25 always has three residues, yet the conformation differs in each of the known structures.

(5) *The new sequence has a VR that is dramatically different from that of all the known structures.* This type is a very long VR, such as the 31 residues at 166–179 in HPH or the 11 residues at 146–151 in thrombin. Similarly, both HPH and thrombin have unusually long loops at 59–62 that contain bound carbohydrate (see Fig. 3, and Tables 3 and 4).

Of the above cases, the first two permit modeling directly or by close analogy to the known structures and are thus readily constructed. Cases 3 and 4 are more difficult to construct from comparative model-building considerations and require some more sophisticated analysis to evaluate the relative merits of the small number of possible conformations for each VR. For the last case, modeling by comparison is impossible. Either additional experimental data are required or much more complex energy analysis is necessary than we can presume will be available in the near future.

For practical modeling purposes, it is important to know how many of the VRs in the sequences examined fall into the respective cases. Modeling based upon Table 4 shows that 47 out of the 86 VRs, or 55%, can be classified as cases 1 or 2 and are thus readily constructed. About 35 or 41% fall into the more difficult cases 3 and 4 and only 4 or $\sim$5% fall into the effectively impossible last class.

A model of each of the new proteins in Figure 3 is constructed by linking together the structural elements, SCRs and VRs, using the most suitable known structure to model each element (Table 4). Considering both the SCRs and the case 1 and 2 VRs, the great majority of each protein can be readily constructed with high confidence in the accuracy and validity of the model.

Although the case 3 and 4 VRs are more difficult to construct, these loops in the known and new structures provide an interesting set of data of varying complexity for developing and testing the efficacy of protein energetics and folding methods, since they require the folding of an independent and end-constrained portion of the molecule (the VR) onto a reasonably well-known structural background that is based upon the SCRs. As a control, some of the loops in the known structures can be built from one of the other known structures. Many of these loops have only a small number of variables and thus many possible conformations can be examined and evaluated in a reasonable amount of time. The additional structural and functional information that would become available from the application of a proper energy analysis to the VRs in a new structure makes the amount of research needed to build these loops very much worthwhile.

Space considerations do not permit the presentation of detailed model structures for each of the eight new sequences (Table 1) nor a discussion of the important functional implications of each of these models. In practice, the reader, armed with Figure 3, Tables 3 and 4, and the modeling methods and results reported here, should be able to construct a model of any of these proteins in a straightforward manner. The accompanying paper (Greer, 1981) presents the detailed model for one of these proteins, blood clotting factor X_a (FAX) together with the implications of that model for the high substrate specificity of this protease, which is critical for the proper functioning of the blood-clotting enzyme cascade.

Several of the members of the mammalian serine protease family, including GSP (Anderson *et al.*, 1978), thrombin (Tsernoglou *et al.*, 1974), and HPH (Hwang, Weiner & Greer, unpublished data), are currently being studied by X-ray diffraction methods. It will be an important test of comparative model building to construct these molecules, both in the SCRs and in the VRs, and then compare the model co-ordinates with the experimentally determined X-ray structures. It will allow a true evaluation of the degree of structural conservation that can be inferred from strong sequence homology, at least in the mammalian serine protease family. It will also determine to what extent the variable regions of the structure can be predicted with reasonable accuracy.

I thank Dr Bruce Bush for his program to calculate solvent accessibility. I also thank Noel Kropf, Christos Tountas, Boris Klebansky and David Yarmush of the Columbia Biology Department Computer Graphics Facility for programs used in the display of these results. This research was supported by National Institutes of Health grant HL-16601 and Facility grant RR-00442 and by the Columbia University Computer Center.

REFERENCES

Anderson, W. F., Matthews, B. W. & Woodbury, R. G. (1978). *Biochemistry*, **17**, 819.

Birktoft, J. J. & Blow, D. M. (1972). *J. Mol. Biol.* **68**, 187–240.

Browne, W. J., North, A. C. T., Phillips, D. C., Brew, K., Vanaman, T. C. & Hill, R. L. (1969). *J. Mol. Biol.* **42**, 65–86.

Dayhoff, M. O. (1972). *Atlas of Protein Sequence and Structure*, vol. 5, National Biomedical Research Foundation, Washington, D.C.

Dayhoff, M. O. (1978). *Atlas of ProteinSequence and Structure*, vol. 5, suppl. 3, National Biomedical Research Foundation, Washington, D.C.

DeHaen, C., Neurath, H. & Teller, D. C. (1975). *J. Mol. Biol.* **92**, 225–259.
Delbaere, L. T. J., Brayer, G. D. & James, M. N. G. (1979). *Nature (London)*, **279**, 165–168.
Fitch, W. M. (1966). *J. Mol. Biol.* **16**, 9–16.
Greer, J. (1980) *Proc. Nat. Acad. Sci., U.S.A.* **77**, 3393–3397.
Greer, J. (1981). *J. Mol. Biol.*
Hartley, B. S. (1970). *Phil. Trans. Roy. Soc. ser. B*, **257**, 77–87.
Huber, R., Kukla, D., Bode, W., Schwager, P., Bartel, K., Diesenhofer, J. & Steigemann, W. (1974). *J. Mol. Biol.* **89**, 73–101.
Jukes, T. H. & Cantor, C. R. (1969). In *Mammalian Protein Metabolism* (Munro, H. N., ed.), pp. 22–132, Academic Press, New York.
Jurasek, L. Olafson, R. W., Johnson, P. & Smillie, L. B. (1976). In *Proteolysis and Physiological Regulation* (Ribbons, D. W. & Brew, K., eds), Miami Winter Symposium, vol. 11, pp. 93–123, Academic Press, New York.
Katayama, K., Ericsson, L. H., Enfield, D. L., Walsh, K. A., Neurath, H., Davie, E. W. & Titani, K. (1979). *Proc. Nat. Acad. Sci., U.S.A.* **76**, 4990–4994.
Kretsinger, R. H. (1976). *Annu. Rev. Biochem.* **45**, 239–266.
Kurosky, A., Barnett, D. R., Lee, T.-H., Touchstone, B., Hay, R. E., Arnott, M. S., Bowman, B. H. & Fitch, W. M. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 3388–3392.
Lee, B. & Richards, F. M. (1971). *J. Mol. Biol.* **55**, 379–400.
Magnusson, S., Sottrup-Jensen, L., Petersen, T. E. & Claeys, H. (1975). In *Boerhaave Symposium on Prothrombin and Related Coagulation Factors* (Hemker, H. C. & Veltkamp, J., eds), pp. 25–46, Leiden University, Leiden.
McLachlan, A. D. (1971). *J. Mol. Biol.* **61**, 409–424.
McLachlan, A. D. (1972). *J. Mol. Biol.* **64**, 417–437.
McLachlan, A. D. & Shotton, D. M. (1971). *Nature New Biol.* **229**, 202–205.
Olafson, R. W., Jurasek, L., Carpenter, M. R. & Smillie, L. B. (1975). *Biochemistry*, **14**, 1168–1177.
Richards, F. M. (1977). *Annu. Rev. Biophys. Bioeng.* **6**, 151–176.
Sawyer, L., Shotton, D. M., Campbell, J. W., Wendel, P. L., Muirhead, H., Watson, H. C., Diamond, R. & Ladner, R. C. (1978). *J. Mol. Biol.* **118**, 137–208.
Shotton, D. M. & Watson, H. C. (1970). *Nature (London)*, **225**, 811–816.
Stroud, R. M., Kay, L. M. & Dickerson, R. E. (1974). *J. Mol. Biol.* **83**, 185–208.
Titani, K., Fujikawa, K., Enfield, D. L., Ericsson, L. H., Walsh, K. A. & Neurath, H. (1975). *Proc. Nat. Acad. Sci., U.S.A.* **72**, 3082–3086.
Tschesche, H., Ehret, W., Godec, G., Hirschauer, C., Kutzbach, C., Schmidt-Kastner, G. & Fiedler, F. (1976). In *Kinins, Pharmacodynamics and Biological Roles* (Sicuteri, F., Back, N. & Haberland, G. L., eds), pp. 123–133, Plenum Press, New York.
Tsernoglou, D., Walz, D. A., McCoy, L. E. & Seegers, W. H. (1974). *J. Biol. Chem.* **249**, 999.
Tulinsky, A., Mani, N. V., Morimoto, C. N. & Vandlen, R. L. (1973). *Acta Crystallogr. sect. B*, **29**, 1309–1322.
Wiman, B. (1977). *Eur. J. Biochem.* **76**, 129–137.
Woodbury, R. G., Katunuma, N., Kobayashi, K., Titani, K. & Neurath, H. (1978). *Biochemistry*, **17**, 811–819.

Edited by A. Klug